electronics

MDPI

Article

Facial Emotion Recognition with Inter-Modality-Attention-Transformer-Based Self-Supervised Learning

Aayushi Chaudhari [1], Chintan Bhatt [2,*], Achyut Krishna [1] and Carlos M. Travieso-González [3]

[1] U & P U. Patel Department of Computer Engineering, Chandubhai S Patel Institute of Technology (CSPIT), CHARUSAT Campus, Charotar University of Science and Technology, Changa 388421, India
[2] Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar 382007, India
[3] Signals and Communications Department, IDeTIC, University of Las Palmas de Gran Canaria, 35001 Las Palmas, Spain
* Correspondence: chintan.bhatt@sot.pdpu.ac.in; Tel.: +91-9909953994

Abstract: Emotion recognition is a very challenging research field due to its complexity, as individual differences in cognitive–emotional cues involve a wide variety of ways, including language, expressions, and speech. If we use video as the input, we can acquire a plethora of data for analyzing human emotions. In this research, we use features derived from separately pretrained self-supervised learning models to combine text, audio (speech), and visual data modalities. The fusion of features and representation is the biggest challenge in multimodal emotion classification research. Because of the large dimensionality of self-supervised learning characteristics, we present a unique transformer and attention-based fusion method for incorporating multimodal self-supervised learning features that achieved an accuracy of 86.40% for multimodal emotion classification.

Keywords: self-attention transformer; multimodality; inter-modality attention transformer; contextual emotion recognition; depth of emotional dimensionality; computer vision; real-time application

check for updates

Citation: Chaudhari, A.; Bhatt, C.; Krishna, A.; Travieso-González, C.M. Facial Emotion Recognition with Inter-Modality-Attention-Transformer-Based Self-Supervised Learning. *Electronics* **2023**, *12*, 288. https://doi.org/10.3390/electronics12020288

Academic Editors: Fabio Mendonca, Morgado Días and Sheikh Shanawaz Mostafa

Received: 16 December 2022
Revised: 30 December 2022
Accepted: 2 January 2023
Published: 5 January 2023

1. Introduction

Emotion recognition and sentiment analysis have recently received a lot of attention due to their numerous applications, such as in human–computer interactions, education, and healthcare robotics. The correlation between the information reflected in and transmitted by the facial expression and the person's contemporaneous emotional state is a hot topic in both customer service and education research. Earlier techniques for encoding data modalities included emotion identification features such as mel-frequency cepstral coefficients (MFCC) [1], elements of facial muscle activity, and glove embedding [2]. Recent studies [3,4] have looked into the use of transfer learning approaches for extracting features from pretrained deep learning (DL) models as opposed to low-level features. The primary purpose of our research was to create contextualized representations from these extracted features using transformer-based architecture, and then use these representations to evaluate low/high degrees of arousal and valence. The goal of our study was to extract face expressions and acoustic sound features [5] from trained DL models for supervised learning. Previous work has combined low-level and deep features rather than representing all modalities using characteristics derived from trained deep learning models [6]. In contrast to earlier research, we have used deep features taken from pretrained self-supervised learning models for representing all input modalities (audio, video, and text) [7–9]. RoBERTa [9], FAb-net [9], and Wav2Vec [9] are three freely accessible pretrained self-supervised learning (SSL) embedded models that we used to represent text, speech, and facial expressions. Emotional indicators that are transferable across speakers, environments, and semantic contents can be incorporated into our representation. Because of the high dimensions of SSL embeddings, the longer SSL feature sequence lengths, and the lack of consistency in the sequence lengths and sizes of SSL embedding features extracted from distinct SSL models across modalities, SSL embedding features provide powerful presentations of all input modalities, but fusing them prior to the final prediction is a difficult task. Although a simple concatenation is a possible solution, additional training parameters are necessary for the embeddings to fully link to the high-dimensional SSL embeddings and increase the likelihood of overfitting the network. We present a novel transformer and attention-based fusion method for incorporating multimodal self-supervised learning features due to the high dimensionality of self-supervised learning characteristics. We also conducted an extensive literature review and discovered that the most recent transformer-based models, such as Fab-Net, Wav2Vec, and RoBERTa, outperform traditional deep learning models. In previous research, we experimented with the machine learning model Support Vector Machine (SVM), where we obtained an accuracy of 61% and with the deep learning model Convolutional Neural Network (CNN), where we obtained an accuracy of 94% on a unimodal approach for emotion classification using image and an accuracy of 82.29% on a unimodal approach for emotion classification using audio [10]. Thus, taking such concerns into account, we developed the reliable and efficient feature fusion approach based on the self-attention transformer. The fusion mechanism was capable of effectively combining input modalities. Two transformers based on self-attention initially enhanced voice and video SSL embeddings. By incorporating a classification (CLS) token that combined the data interwoven throughout the entire sequence, this phase altered both speech and video sequences; it added details from the other modality to the sequence representation in each modality. We used modality-specific CLS tokens in this phase. In the concluding section, we developed a method based on the Hadamard product to identify the salient characteristics of each modality. In conclusion, our primary contributions appeared to be multimodal emotion classification, where we leveraged trimodal SSL features derived from three separately pretrained SSL architectures. We provided an innovative transformer-based fusion technique for fusing SSL characteristics with variable embeddings, sizes, and sequence lengths.

2. State of the Art

Computer researchers are looking into a variety of techniques for replicating techniques that can successfully build effective algorithms in order to create analytical models for face-image analyses that can successfully identify human expressions. Computer vision, artificial intelligence, and pattern recognition are three fields of study that have a history of developing novel solutions to emotional perception problems. Most of the earlier research in this area focused on gathering data from unimodal systems. Machines that can only read facial gestures [11] or speech sounds [12] have used these to predict emotion. After thorough research over time, multimodal systems that predict mood using various inputs have ultimately showed to be more accurate and efficient. The designing of convolutional neural networks (CNNs), a form of neural network, over the past few years has led to another significant milestone. When a signal that is inputted is broken down (de-convolved) into a collection of invariant features, CNNs have been used as general problem solvers, offering reliable ways to extract pertinent features (like texture, key-points, and corners). Numerous research articles, such as [13–15], have provided a very thorough explanation of facial expression or emotion identification systems. They have shed information on numerous facets of picture/video capturing, after-processing, feature extraction, and recognition, as well as readily available datasets, which are crucial inputs for effectively training an emotion recognition system. Other research papers have expanded on the joint application of facial emotion recognition (FER) and subordinate sources, including those on the expression of several facial emotions [16], utilizing augmented reality [17], sentiment categorization, along with gender identification facial expression recognition [18], and an evaluation of techniques particularly designed for real against produced emotional face appearances [19]. At the moment, single-medium emotion recognition, including text, speech, and image, is the focus of most emotion recognition research [20]. Multimodal emotion identification systems fully account for the interplay between audio, text, image, and other modalities [21]. A multimodal fusion approach for voice expressions was presented by Dong Liu et al. Speech expressions use convolutional networks with long short-term memory (LSTM) cells to capture the correlation between and within the various modalities, while video-based expressions use the Inception-ResNet-v2 network to extract the feature data [22]. Comparing multimodal approaches to unimodal emotion identification techniques, the robustness of the emotion recognition system was improved [23]. In this study, multimodal emotion recognition typically covered modalities including speech, text, image, video, and many more. The Gaussian mixed model (GMM) and hidden Markov model (HMM) have both been used extensively in studies to extract facial features [24]. The authors suggested scale-invariant textures, Gaussian mixture features, and multiresolution curvelet-transform-based symbiosis for image retrieval and classification [25]. CNNs serve as the foundational model of deep learning, which has been chosen in conjunction with other network models in emotion recognition, including LSTM and recurrent neural networks (RNNs). These optimization models have produced great results in the domain of emotion recognition [26]. For recognizing speech emotions, [27] suggested a parallel convolutional RNN incorporating spectral characteristics. To categorize emotions using the SoftMax classifier, several high-level characteristics were merged and then batch-normalized, improving emotion detection. By combining Wavegram-Logmel features, which combine the wavegram and log-mel spectrograms and speech characteristics in the waveform, which are directly retrieved from one-dimensional time–domain waveforms, Hussain et al. developed a revolutionary method. By combining the traits from each previously trained SSL embedding model with three distinct modalities—video, audio, and text—Shamane et al. used SSL to illustrate emotion recognition using a multimodal approach. Here, the author combined multimodal SSL features by using innovative transformers and attention-based fusion mechanisms [28]. The three separate modes that Baijun et al. considered are sound, text, and video. These modalities were structured and optimized on the Multimodal EmotionLines Dataset (MELD). To determine the emotion, the author combined the EmbraceNet architecture with a transformer-based cross-modality fusion [29]. Tzirakis et al. presented a multimodal approach for emotion recognition using CNNs for extracting audio features and ResNet (50 layers) for extracting features from images. Several authors have used LSTM for working upon outliers for modeling the context [30]. Kansizoglou et al. used the emotion identification aspect to recognize the personality of humans by tracking the continuous variations of emotions of an individual. The author proposed a system which used face landmarks to train a deep recurrent neural network and estimated two coefficients of emotions that were arousal and valence [31]. Zhang at el. proposed an audio–visual spatiotemporal deep neural network that contained a visual block containing a 2D-CNN pretrained temporal convolutional network, an aural block containing a parallel temporal convolutional network (TCN), and a leader–follower attentive fusion block, which combined the audio–visual information [32]. Kansizoglou et al. developed a method for an artificial agent to conceive of and eventually understand the personality of a human by observing his/her emotional fluctuations over the course of his/her encounter. To accomplish this, the subject’s facial landmarks were retrieved and fed into deep neural network architecture that calculated the two coefficients of human emotions, namely arousal and valence [33].

3. Datasets

RAVDESS Audio–Visual Dataset

Twenty-four professional actors were involved in creating this dataset with a neutral North American accent, representing a gender-balanced representation of the population, and performing lexically matched sentences in the database. Both speech and music can display a range of emotions, including fear, surprise, and disgust [34]. There were two emotional intensity levels for each expression, in addition to a neutral expression. Various forms of emotion for face-and-voice, face-only, and voice-only were all available for all situations. Each of the 7356 recordings in the collection received 10 ratings for emotional validity, intensity, and sincerity. Two hundred and forty-seven people who represented the untrained research participants from North America provided their ratings. Seventy-two additional participants supplied test-retest information. There were three format options for each actor's recorded performance: video-only (VO), audio–visual (AV), and audio-only (AO). We considered eight different emotion categories to identify human emotion, namely Calm, Sad, Happy, Neutral, Surprised, Disgust, Fearful, and Angry. Figure 1 represents the sample images from the RAVDESS Audio–Visual Dataset for various emotion categories.

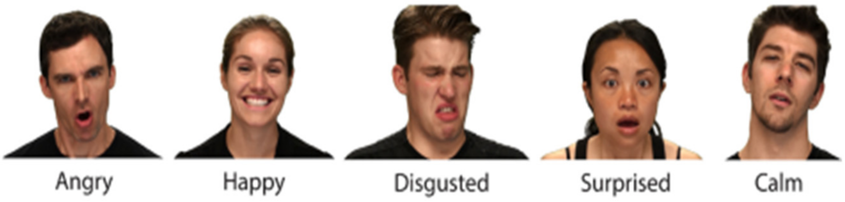

Figure 1. RAVDESS dataset specimens of emotion representations.

4. Methodology

This research aimed to develop an automated emotion identification system that successfully integrates all input feature modalities with SSL properties to represent important information that may be obtained through in-person mental state monitoring. A computer system can provide an alarm when someone displays fear, scorn, misery, or any other comparable emotion. This can assist medical professionals in monitoring a patient's mental state. If anything goes wrong, professionals can alter their strategy and technique for treating the patient.

We created a fusion method that could be easily expanded to incorporate SSL features from other modalities. Hadamard computation for information extraction, modality-specific CLS token-based inter-modality attention (IMA), and embedding modifications using CLS tokens for emotion categorization were the unique elements of our transformer-based fusion approach. Finally, we explained the feature selection based on Hadamard computation.

4.1. Multimodal Feature Extraction Using SSL Models

We used three pretrained SSL models in this study. All model checkpoints were derived from openly accessible sources. The FAIRSEQ codebase [35] was utilized to gain access to the pretrained RoBERTa [9] and Wav2Vec [9] models as well as to extract text and audio SSL features. Using the publication [8], we could retrieve the pretrained Fab-Net model to extract video modality features. By using the Retina Face [36] facial recognition system, we clipped faces from each video frame to extract attributes from movies. Fab-Net's pretrained model was then utilized to extract features from the video stream for each frame that featured a face. We did not enhance any SSL models using the multimodal emotion identification datasets. The frozen SSL model was used to retrieve the features for each data modality. Each modality had different SSL properties in terms of size and maximal training sequence lengths, as stated in Table 1.

Table 1. SSL model representation and the respective embedding size and sequence length.

Model	Embedding Size	Max Sequence Length
Wav2Vec	512	935
Fab-Net	256	300
RoBERTa	1024	512

(1). Wav2Vec

Layers of temporal convolution were used in the construction of the Wav2Vec [37] architecture, and the self-supervised training pre-text task used the contrastive predictive coding [38] concept.

Figure 2 represents the unprocessed audio waveform that can be represented by the context representation C. A maximum audio shape duration of 9.5 s and a maximum embedding size of 512 were taken into consideration. The network, which had 35 M parameters, was pretrained using audio from the LibriSpeech collection totaling 960 h [39]. The FAIRSEQ repository was used to obtain the pretrained model checkpoint [35].

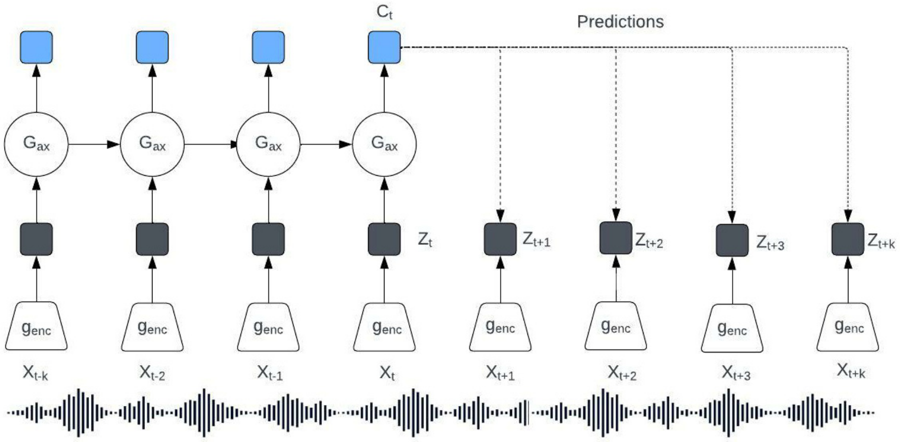

Figure 2. Generation architecture for Wav2Vec.

(2). Fab-Net

We produced embeddings using Fab-Net [8] that had already been trained in every picture in frame which featured the face of the speaker. The Fab-Net pre-text challenge was specifically created to motivate the network to learn facial characteristics representing locations, poses, and expressions. The network was expected to transform the input source frames into the destination frames by predicting the flow field in between them, since only the extracted features matched source and target frames. The network had to approach the offset that must take place in the pixels of the input images in order to produce the target result image. This delicate action required a network to extract the data needed to calculate the flow field, including head location and expression into the source and destination embeddings. Despite being taken at different angles and positions, the input and destination image frames both came from identical face tracks of the same person. Using two substantial defocused datasets, the network was pretrained [40]. The embedding dimension was 256.

4.2. Feature Extraction Mechanism

Three SSL pretrained models allowed us to extract characteristics from various forms of raw data; SSL features come in a variety of sizes and maximal training sequence lengths depending on the modality. To extract speech and text SSL features, the FAIRSEQ source [35] was used to obtain the pretrained models RoBERTa [9] and Wav2Vec [9]. We obtained the pretrained Fab-Net model from the publication [41] and extracted the features for the video modality. In order to extract characteristics from movies, we cut faces out of each frame of a video using the Retina Face [42] face identification algorithm. Later, for every frame which had a human face in it, we utilized the previously trained Fab-Net model to record video modality information.

4.3. SSL Embedding

We attempted to develop a technique that can explain a long embedding sequence connected with different modalities using only one embedding. For this purpose, we applied self-attention to every embedding sequence and added a trainable vector called CLS to the Wav2Vec and Fab-Net embeddings (A and V).

We chose an initial unique token named CLS for our embedding sequence adaptation period for classification because the way that BERT [42] or RoBERTa [37] models express a full sequence inspired us. Since the self-attention mechanism is bidirectional in BERT-based models (past and future), the CLS token in the series was encoded with the data to its right. As a result, as a compact representation, the CLS token may be used to address classification difficulties such as emotion recognition. Given that Wav2Vec and Fab-Net embedding sequences do not share a structure with BERT, we simply prepended CLS tokens to them in our model. We selected the text embedding sequence, as RoBERTa works as a normal BERT-based model. Thus, we were able to compute inter-modality attention (IMA) more quickly and create a straightforward late fusion mechanism due to the availability of three CLS tokens that represented three modalities.

Figure 3 depicts the sequence for speech and video embedding modification using two transformer blocks. SSL model features have a huge embedding size and a lengthy sequential length.

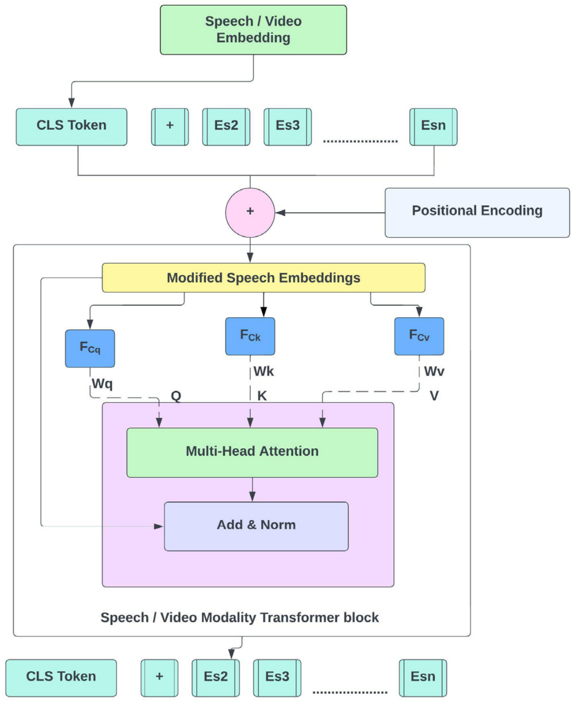

Figure 3. Transforming speech and video during the embedding process.

4.4. IMA-Based Fusion Layer

The inter-modality attention (IMA) layer worked similar to the self-attention layer, with the exception that it created the key (K) and value (V) vectors from the embedding sequence of the other modality and the query (Q) vector from the CLS token of the other modality. Three embedding sequences were provided to the inter-modality attention fusion layer as inputs, with the CLS token appearing as the first token in each sequence. Since the CLS token of each modality collected information from the sequence, the inter-modality attention was roughly split between the whole embedding sequence of one modality and its CLS token. Thus, there were six inter-modality-attention-based transformer blocks with a Q vector formed from a modality's CLS token and K-V vector derived from the modality's entire embedding sequence.

Figure 4 showcases the process of using multi-head-attention- and inter-modality-attention-based approaches for classifying emotion. On a collection of inquiries that were arranged into the Q matrix, we computed the attention function. The matrices K and V also included both the keys and values. The sign dk simultaneously indicates the query, value, key, and dimensionality of the query vector. The output matrix is calculated as follows:

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}\left(\frac{Qk^{t}}{\sqrt{d_k}}\right)\mathrm{V} \tag{1}$$

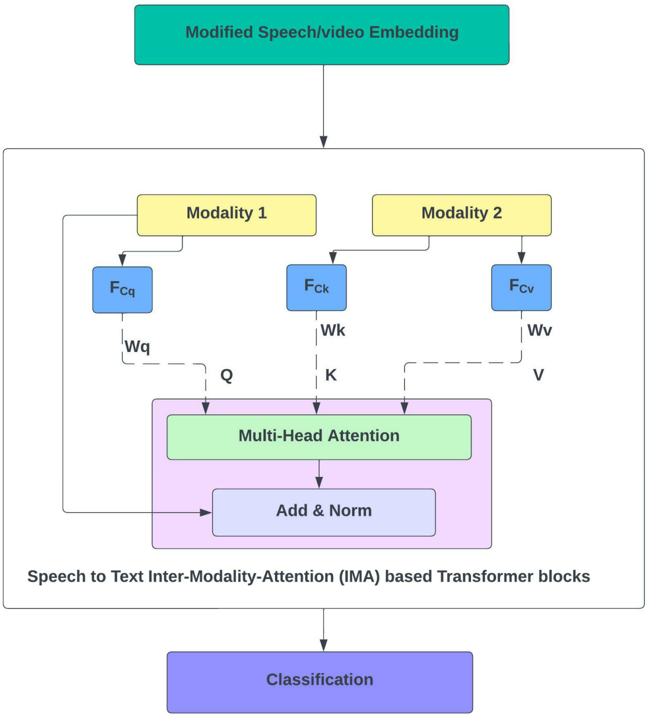
Modified Speech/video Embedding
Modality 1
Modality 2
Fcq
Fck
Fcv
Wq
Wk
Wv
Q
K
V
Multi-Head Attention
Add & Norm
Speech to Text Inter-Modality-Attention (IMA) based Transformer blocks
Classification

Figure 4. The process of classifying the emotions using multi-head-attention- and inter-modality-attention-based transformers.

4.5. Hadamard Product

Before the prediction layer, we looked at potential combinations. The obvious technique for aggregating information appeared to be the concatenation of six tokens. However, before concatenating in our work, we further simplified the CLS token. Based on the fundamental mode of the Q vector used in inter-modality attention computation, the three pairs of the six CLS embeddings were categorized. The key information produced by the video, speech (audio), and text modalities was extracted using the Hadamard product between six CLS tokens, which were retrieved from the inter-modality attention layer by computing the Hadamard product (*) among the CLS token pairs that corresponded to the same core modality. Vfinal, Afinal, and Tfinal are the three vectors that are produced following the computation of the Hadamard product between the outputs of the same core inter-modality attention modality. It was possible to easily extract the shared data among the two CLS models using the Hadamard product. After the Hadamard calculation, the final three forms were concatenated and sent via a prediction layer.

4.6. Synopsis of the Fusion Method

Our approach for the integrated, self-supervised fusion transformer of the presented fusion model used the self-supervised embedding of the three modalities as inputs. Before applying a self-attention transformer, speech and visual modalities were first given a special CLS token. These CLS tokens for each modality were used to aggregate the data from the full series. Transformer blocks were used to modify Wav2Vec's and Fab-Net's self-supervised embeddings. Each embedding cycle that covered both text and speech modalities was first inserted using different tokens from CLS speech and video. The two distinct blocks of self-attention-based transformers were then independently passed through with altered embedding sequences. Figure 5 represents the in-short architecture of emotion classification using a Hadamard product and inter-modality attention transformer. Each inter-modality attention transformer took as inputs the entirety of a modality's embedding sequence in addition to the CLS token from that modality. For example, the notation "Audio Text" denoted that the CLS token belonged to the voice modality whilst the other embedding sequence belonged to the text modality. Here, the shift from speech to text was highlighted. Every single inter-modality attention transformer generated a CLS token that was enhanced by embedding-level knowledge acquired from the opposite modality. Six transformer models based on inter-modality attention (IMA) made up the fusion process. A CLS token from one modality could access the entire embedding sequence of another modality via each of the transformer stages and gather crucial cross-modal data. Finally, the fusion process generated six CLS tokens improved using intermodal data.

$$\begin{aligned} &Hadamard\ Product\ of\ final\ fusion \\ &\quad = \Big[(Video \rightarrow Speech)_{[cls]} \odot (Video \rightarrow Text)_{[cls]} \\ &\quad \odot (Speech \rightarrow Video)_{[cls]} \odot (Speech \rightarrow Text)_{[cls]} \\ &\quad \odot (Text \rightarrow Video)_{[cls]} \odot (Text \rightarrow Speech)_{[cls]} \end{aligned} \tag{2}$$

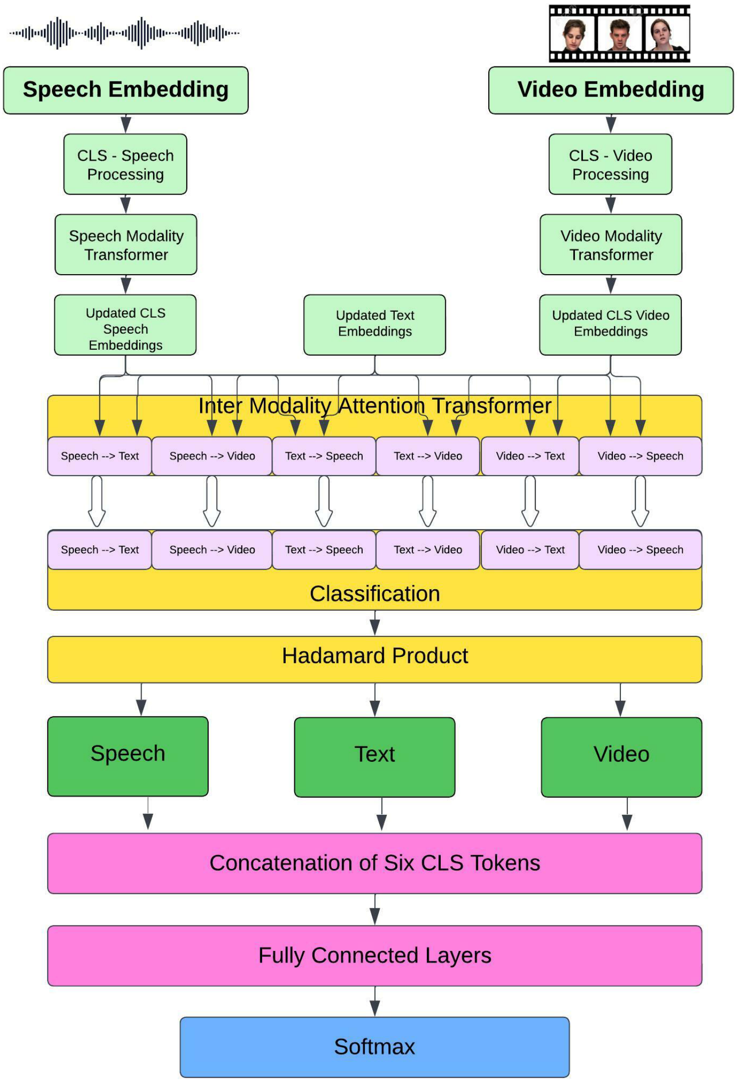

Figure 5. Hadamard product voice and video embedding for emotion categorization from start to finish.

Prior to the concatenation layer, the Hadamard product was employed to extract significant features from CLS token sets that correlated to the same modality. To obtain the output probabilities, the combined vectors were then passed through a fully connected layer.

4.7. Implementation Details

This section thoroughly describes the model's application and the experimental setting. To put our idea into action, we used FAIRSEQ [35]. The training was carried out on a machine with 128 GB of RAM and a Nvidia Tesla v100 with a storage capacity of 20 TB, 32 GB of RAM, and a 576-core Intel Xeon 2 processor.

5. Results

The objective of this research was to build a fusion mechanism that would use SSL properties to represent all input feature modalities. The major objective of our work was to develop a process for fusion that could be simply enhanced with self-supervised learning characteristics from some multimodal emotion recognition techniques. We thoroughly investigated how features of self-supervised learning could depict all modalities while resolving the issues of the SSL features' high dimensionality. A variety of data-stream pretrained self-supervised learning models are now available to the open-source community as a result of the self-supervised learning (SSL) paradigm's ability to operate with publicly available unlabeled data. Six inter-modality-attention-based transformers and two self-supervised-attention-based transformers constituted the majority of the fusion process we suggest using. The fusion was primarily produced to deal with discrete self-supervised sequence embedding while carefully taking into account variations in self-supervised embeddings produced by architectures of various pretrained algorithms that the academic community can acquire without effort. There is an increasing number of self-supervised learning models that have been pretrained for various streams because the self-supervised learning (SSL) paradigm can make use of easily accessible unlabeled data. The core of our proposed fusion technique consisted of two self-supervised-attention-based transformers and six inter-modality-attention-based transformers. Our primary goal of the fusion technique was to handle discrete self-supervised learning (SSL) embedding sequences while taking into account variations in self-supervised sequence embeddings generated by other architectures that are pretrained. This cause was the basis for the tests demonstrating the self-supervised learning (SSL) features' potential to enhance performance for the multimodal emotion recognition task utilizing publicly accessible RAVDESS datasets. The results are shown in the table below. Due to these factors, studies using publicly accessible RAVDESS datasets have demonstrated self-supervised learning (SSL) features' superior performance when used for multimodal emotion identification. The findings are shown in Table 2.

In one or more dimensions of space, the dimensional model depicted the emotional states. Wilhelm Max Wundt classified human emotions into three groups in 1897: "pleasurable vs. unpleasurable", "arousing vs. subduing", and "strain vs. relaxation" [43]. Harold Schlosberg first proposed the three qualities of "pleasantness-unpleasantness", "attention-rejection", and "degree of activation" in 1954 [43]. Most dimensional models divide emotional states into the "valence" and "arousal" dimensions. The figure below shows the mean ratings for arousal and valence in the audio-only (n = 21), visual-only (n = 14), and audio–visual (n = 8) conditions. From Table 2, we recognized that our proposed approach could classify all the eight emotions precisely and it worked best on the Happy, Fearful, and Surprised emotions.

Figure 6 represents the plot of the confusion matrix for eight emotion categories that included sad, calm, fear, surprise, angry, happy, neutral, and disgust.

Figures 7 and 8 represent the plotting of training and testing accuracy and loss with regard to every epoch. This graph also showcases that there were no over-fitting or unbefitting in our training method.

Table 2. Precision, recall, and accuracy metrics per emotion for the top model that achieved an accuracy of 86.40%.

Emotions	**Modalities**											
	Wav2Vec			**Fab-Net**			**RoBERTa**			**Multimodal**		
	PRE	**F1**	**REC**	**PRE**	**F1**	**REC**	**PRE**	**F1**	**REC**	**PRE**	**F1**	**REC**
Neutral	0.84	0.75	0.85	0.83	0.77	0.82	0.86	0.80	0.86	86.31	77.33	70.25
Calm	0.85	0.80	0.78	0.87	0.83	0.79	0.84	0.82	0.81	81.45	81.66	75.63
Happy	0.88	0.78	0.71	0.85	0.80	0.80	0.87	0.79	0.77	89.79	79.10	79.25
Sad	0.86	0.79	0.99	0.84	0.80	0.93	0.85	0.75	0.89	81.22	78.28	81.88
Angry	0.82	0.81	0.99	0.85	0.88	0.90	0.83	0.80	0.80	81.78	83.00	82.37
Fearful	0.78	0.80	0.88	0.85	0.79	0.81	0.80	0.83	0.88	88.26	80.66	83.62
Disgusted	0.85	0.85	0.96	0.86	0.86	0.80	0.81	0.84	0.95	83.99	85.12	85.50
Surprised	0.87	0.89	0.98	0.82	0.90	0.91	0.85	0.86	0.98	89.50	88.33	77.87
Avg Weightage	84.37	80.87	89.25	84.62	82.87	84.50	83.87	81.12	86.75	85.28	81.68	79.54

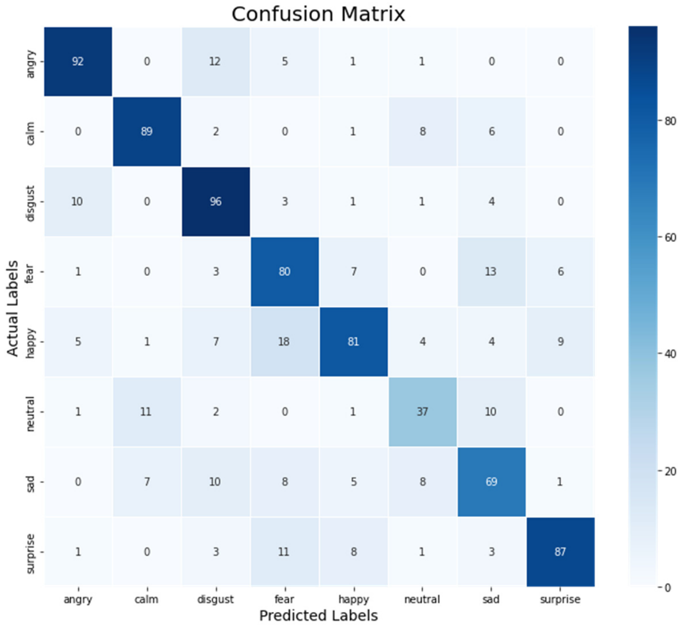

Figure 6. Confusion matrix for all emotion categories.

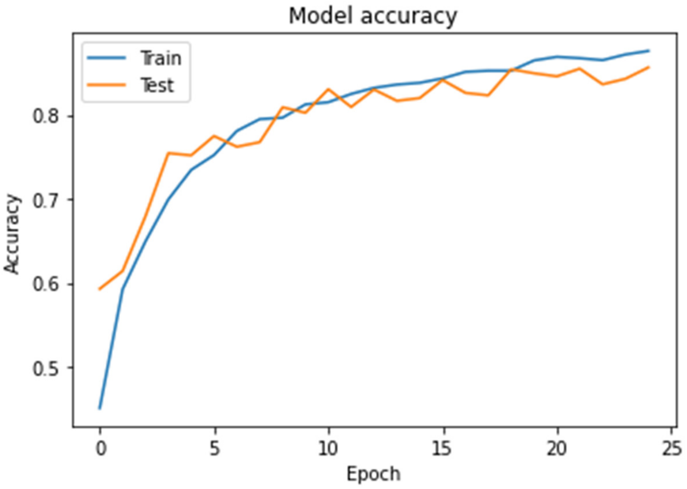

Figure 7. Plot of training and testing accuracy with regard to each epoch.

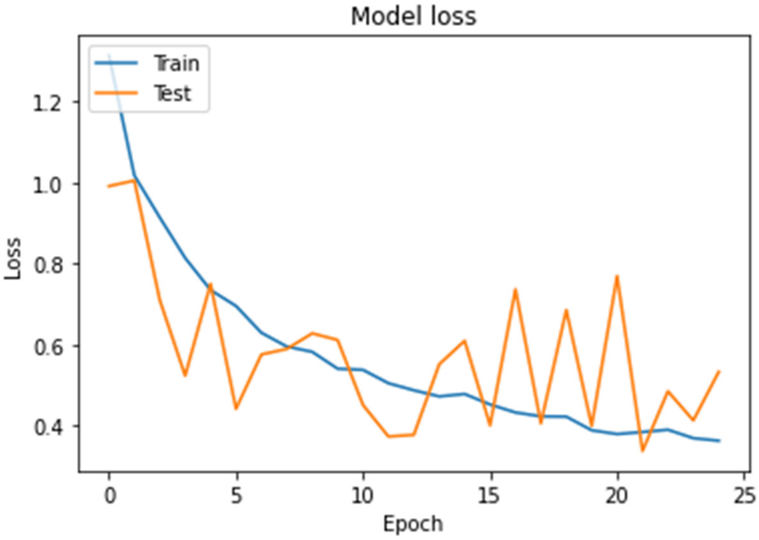

Figure 8. Plot of training and testing loss with respect to each epoch.

Figure 9 shows the graphical representation of the results for all emotion categories in two-dimensional form taking arousal and valence into consideration. We have considered recognizing emotion categories into two forms (i.e., arousal and valence for identifying acute feelings of the person).

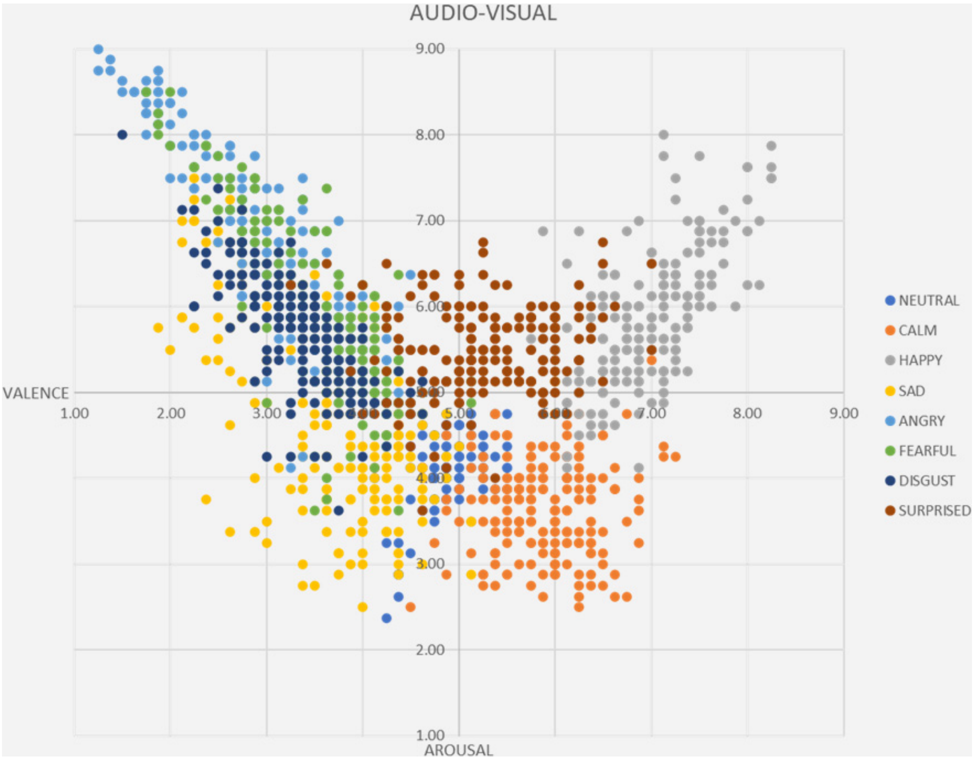

Figure 9. Emotion classification of eight categories using valence and arousal in two-dimensional space.

6. Comparison with Existing Studies

The comparative analysis of the multimodal approach for the suggested mode and existing study is shown in Table 3. We carried out a series of comparisons and selected the latest work in order to combine the modalities of audio, face, and text. Siriwardhana et al. [28] created a multimodal fusion model and attained cutting-edge outcomes. By combining numerous modalities as opposed to a single modality, the accuracy and F1-scores were enhanced, according to the findings of the earlier investigations. Similarly, Xie et al. [29] presented multimodal as well as unimodal approaches for recognizing emotions for both audio and face modalities. We suggest a method for identifying emotions and determining their valence and arousal when compared with the following studies.

- Multi-level Multi-Head Fusion Attention RNN Model [28]

The authors presented a novel transformer-based fusion mechanism for fusing SSL characteristics with arbitrarily embedded embeddings, sizes, and sequence durations. They further tested and compared their model's robustness and generalizability on four publicly available multimodal datasets. Their fusion mechanism, called the "Self-Supervised Embedding Fusion Transformer (SSE-FT)," used tri-modal SSL features extracted from three independently pretrained SSL architectures for multimodal emotion recognition. Their framework consisted mainly of two self-attention-based transformers and six inter-modality-attention (IMA)-based transformer blocks. The speech and video SSL embeddings were first modified by two self-attention-based transformers. This stage modified both speech and video recordings by adding a special token called CLS that could consolidate the information inherent in the entire sequence. Six IMA-based transformers then processed the three SSL embedding sequences, enhancing each modality sequence representations with information from another modality. They used CLS tokens particular to each modality and then employed a Hadamard-product-based algorithm to determine the most important characteristic in each modality.

- Robust Cross-modality Fusion [29]

This study demonstrated robust multimodal emotion categorization architecture that combined three independent input modalities via cross-modal transformer fusion. The architecture took into account both the modalities' joint relationships and the robust fusion of different representation vector sources. Three distinct prediction methods were adapted to discern emotion using audio, video, and text-based inputs. GPT, WaveRNN, and FaceNet+GRU were used to train text, audio, and image inputs, respectively. The proposed transformer-based fusion method demonstrated its capacity to perform the job of multimodal feature fusion from several pretrained models using EmbraceNet. The attention results from the cross-modal systems were employed by EmbraceNet to build a fused depiction of the emotion embedding vectors.

Table 3. Comparative study of proposed model with existing work on multimodal approaches.

Experiments	Modality	Accuracy	F1-Score
Multi-level Multi-Head Fusion Attention RNN Model [28]	Multimodal (Audio + Facial + Text)	64.3	63.9
Robust Cross-Modality Fusion [29]	Audio	48.4	32.1
	Facial	47.8	31.4
	Text	62.6	61.2
	Multimodal (Audio + Facial + Text)	65.0	64.0
Our Proposed Approach	Multimodal (Audio + Facial + Text)	87.6	81.68

Our suggested method achieved greater accuracy and equivalent F1-scores, which were on par with state-of-the-art performance and demonstrated the robustness in multimodal emotion categorization.

7. Conclusions

We proposed an approach that used unsupervised data that is extensively available using self-supervised learning (SSL) algorithms to recognize the emotion. By utilizing this strategy, we were able to save time in retraining the model or starting from scratch and to utilize pretrained self-supervised learning algorithms that are currently available. Using self-supervised learning as an input indicated that the features generated had high dimensions and were regarded as high-level features that required a trustworthy and in-depth fusion process. The outcomes indicated that we can successfully tackle the problem of multimodal emotion identification using the self-supervised learning (SSL) and inter-modality interaction approaches. By using pretrained self-supervised learning algorithms for the extraction of features, we focused on improving the task of emotion identification. We developed a multimodal fusion technique that was a transformer-based method to achieve our goal. Moreover, we acutely identified our emotion categories by applying them in two dimensions (i.e., arousal and valence). Initially, we demonstrated that our technique could outperform earlier state-of-the-art methods by comparing our model to strong baselines from RAVDESS datasets. In the future, we hope to experiment with recognizing emotions from contextual data and categorizing them into three dimensions: arousal, valence, and dominance. We also intend to put our model to the test in the medical arena to assist specialists in accurately diagnosing patients.

Author Contributions: Conceptualization, A.C.; Methodology, A.C. and A.K.; Data curation, A.K.; Writing—original draft, A.C.; Writing—review & editing, C.M.T.-G.; Supervision, C.B. and C.M.T.-G.; Project administration, C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is available in a publicly accessible repository. Publicly available datasets were analyzed in this study. This data can be found on: Livingstone SR, Russo FA (2018), The Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English, and PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kansizoglou, I.; Bampis, L.; Gasteratos, A. An Active Learning Paradigm for Online Audio-Visual Emotion Recognition. *IEEE Trans. Affect. Comput.* **2019**, *13*, 756–768. [CrossRef]
2. Yoon, S.; Byun, S.; Jung, K. Multimodal Speech Emotion Recognition Using Audio and Text. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018.
3. Han, Z.; Zhao, H.; Wang, R. Transfer Learning for Speech Emotion Recognition. In Proceedings of the 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), Washington, DC, USA, 27–29 May 2019; pp. 96–99.
4. Ezzeldin, M.; ElShaer, A.; Wisdom, S.; Mishra, T. Transfer learning from sound representations for anger detection in speech. *arXiv* **2019**, arXiv:arXiv.1902.02120. [CrossRef]
5. Nagarajan, B.; Oruganti, V.R.M. Deep net features for complex emotion recognition. *arXiv* **2018**, arXiv:1811.00003. [CrossRef]
6. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *arXiv* **2019**, arXiv:arXiv.1911.05544. [CrossRef]
7. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *arXiv* **2019**, arXiv:arXiv.1902.06162. [CrossRef]
8. Wiles, O.; Koepke, A.S.; Zisserman, A. Self-supervised learning of a facial attribute embedding from video. *arXiv* **2018**, arXiv:1808.06882.
9. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. Wav2vec: Unsupervised pre-training for speech recognition. *arXiv* **2019**, arXiv:arXiv.1904.05862. [CrossRef]
10. Chaudhari, A.; Bhatt, C.; Krishna, A.; Mazzeo, P.L. ViTFER: Facial Emotion Recognition with Vision Transformers. *Appl. Syst. Innov.* **2022**, *5*, 80. [CrossRef]
11. Levi, G.; Hassner, T. *Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns*; SC/Information Sciences Institute, the Open University of Israel: Marina del Rey, CA, USA, 2014.
12. Han, K.; Yu, D.; Tashev, I. *Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine*; Department of Computer Science and Engineering, The Ohio State University: Columbus, OH, USA; Microsoft Research, One Microsoft Way: Redmond, WA, USA, 2014.
13. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [CrossRef]
14. Huang, Y.; Chen, F.; Lv, S.; Wang, X. Facial Expression Recognition: A Survey. *Symmetry* **2019**, *11*, 1189. [CrossRef]

15. Dhwani, M.; Siddiqui, M.F.H.; Javaid, A.Y. Facial emotion recognition: A survey and real-world user experiences in mixed reality. *Sensors* **2018**, *18*, 416.
16. Ullah, S.; Tian, W. A systematic literature review of recognition of compound facial expression of emotions. In Proceedings of the ICVIP 2020: 2020 the 4th International Conference on Video and Image Processing, Xi'an, China, 25–27 December 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 116–121. [CrossRef]
17. Rajan, S.; Chenniappan, P.; Devaraj, S.; Madian, N. Facial expression recognition techniques: A comprehensive survey. *IET Image Process.* **2019**, *13*, 1031–1040. [CrossRef]
18. Gupta, A.; Sharma, D.; Sharma, S.; Agarwal, A. Survey paper on gender and emotion classification using facial expression detection. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020, Delhi, India, 20–22 February 2020. [CrossRef]
19. Jia, S.; Wang, S.; Hu, C.; Webster, P.J.; Li, X. Detection of genuine and posed facial expressions of emotion: Databases and methods. *Front. Psychol.* **2021**, *11*, 3818. [CrossRef]
20. Rao, K.P.; Chandra, M.V.P.; Rao, S. Assessment of students' comprehension using multi-modal emotion recognition in e-learning environments. *J. Adv. Res. Dyn. Control Syst.* **2019**, *10*, 767–773.
21. Huddar, M.G.; Sannakki, S.S.; Rajpurohit, V.S. Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. *Int. J.Multimed. Inform. Retriev.* **2020**, *9*, 103–112. [CrossRef]
22. Liu, D.; Wang, Z.; Wang, L.; Chen, L. Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning. *Front. Neurorobot.* **2021**, *15*, 697634. [CrossRef]
23. Elleuch, H.; Wali, A. Unwearable multi-modal gestures recognition system for interaction with mobile devices in unexpected situations. *IIUM Eng. J.* **2019**, *20*, 142–162. [CrossRef]
24. Andy, C.; Kumar, S. An appraisal on speech and emotion recognition technologies based on machine learning. *Int. J. Automot. Technol.* **2020**, *8*, 2266–2276. [CrossRef]
25. Engin, M.A.; Cavusoglu, B. Rotation invariant curvelet based image retrieval and classification via Gaussian mixture model and co-occurrence features. *Multimed. Tools Appl.* **2019**, *78*, 6581–6605. [CrossRef]
26. Liu, X.; Zhou, F. Improved curriculum learning using SSM for facial expression recognition. *Vis. Comput.* **2020**, *36*, 1635–1649. [CrossRef]
27. Jiang, P.; Fu, H.; Tao, H.; Lei, P.; Zhao, L. Parallelized Convolutional Recurrent Neural Network with Spectral Features for Speech Emotion Recognition. *IEEE Access* **2019**, *7*, 90368–90377. [CrossRef]
28. Siriwardhana, S.; Kaluarachchi, T.; Billinghurst, M.; Nanayakkara, S. Multimodal Emotion Recognition with Transformer-Based Self Supervised Feature Fusion. *IEEE Access* **2020**, *8*, 176274–176285. [CrossRef]
29. Xie, B.; Sidulova, M.; Park, C.H. Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Cross-modality Fusion. *Sensors* **2021**, *21*, 4913. [CrossRef]
30. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [CrossRef]
31. Ioannis, K.; Misirlis, E.; Tsintotas, K.; Gasteratos, A. Continuous Emotion Recognition for Long-Term Behavior Modeling through Recurrent Neural Networks. *Technologies* **2022**, *10*, 59.
32. Zhang, S.; Ding, Y.; Wei, Z.; Guan, C. Continuous emotion recognition with audio-visual leader-follower attentive fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
33. Kansizoglou, I.; Misirlis, E.; Gasteratos, A. Learning Long-Term Behavior through Continuous Emotion Estimation. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference, Corfu, Greece, 29 June–2 July 2021.
34. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
35. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019.
36. Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-stage Dense Face Localisation in the Wild. *arXiv* **2019**, arXiv:1905.00641.
37. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
38. Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
39. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210. [CrossRef]
40. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018.
41. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019.

42. Wundt, W.M.; Judd, C.H. *Outlines of Psychology (Vol. 1)*; Scholarly Press: Cambridge, MA, USA, 1897.
43. Schlosberg, H. Three dimensions of emotion. *Psychol. Rev.* **1954**, *61*, 81. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.